

MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark

Year: 2024 | Citations: 979 | Authors: Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra

Abstract

In the age of large-scale language models, benchmarks like the Massive Multitask Language Understanding (MMLU) have been pivotal in pushing the boundaries of what AI can achieve in language comprehension and reasoning across diverse domains. However, as models continue to improve, their performance on these benchmarks has begun to plateau, making it increasingly difficult to discern differences in model capabilities. This paper introduces MMLU-Pro, an enhanced dataset designed to extend the mostly knowledge-driven MMLU benchmark by integrating more challenging, reasoning-focused questions and expanding the choice set from four to ten options. Additionally, MMLU-Pro eliminates the trivial and noisy questions in MMLU. Our experimental results show that MMLU-Pro not only raises the challenge, causing a significant drop in accuracy by 16% to 33% compared to MMLU but also demonstrates greater stability under varying prompts. With 24 different prompt styles tested, the sensitivity of model scores to prompt variations decreased from 4-5% in MMLU to just 2% in MMLU-Pro. Additionally, we found that models utilizing Chain of Thought (CoT) reasoning achieved better performance on MMLU-Pro compared to direct answering, which is in stark contrast to the findings on the original MMLU, indicating that MMLU-Pro includes more complex reasoning questions. Our assessments confirm that MMLU-Pro is a more discriminative benchmark to better track progress in the field.